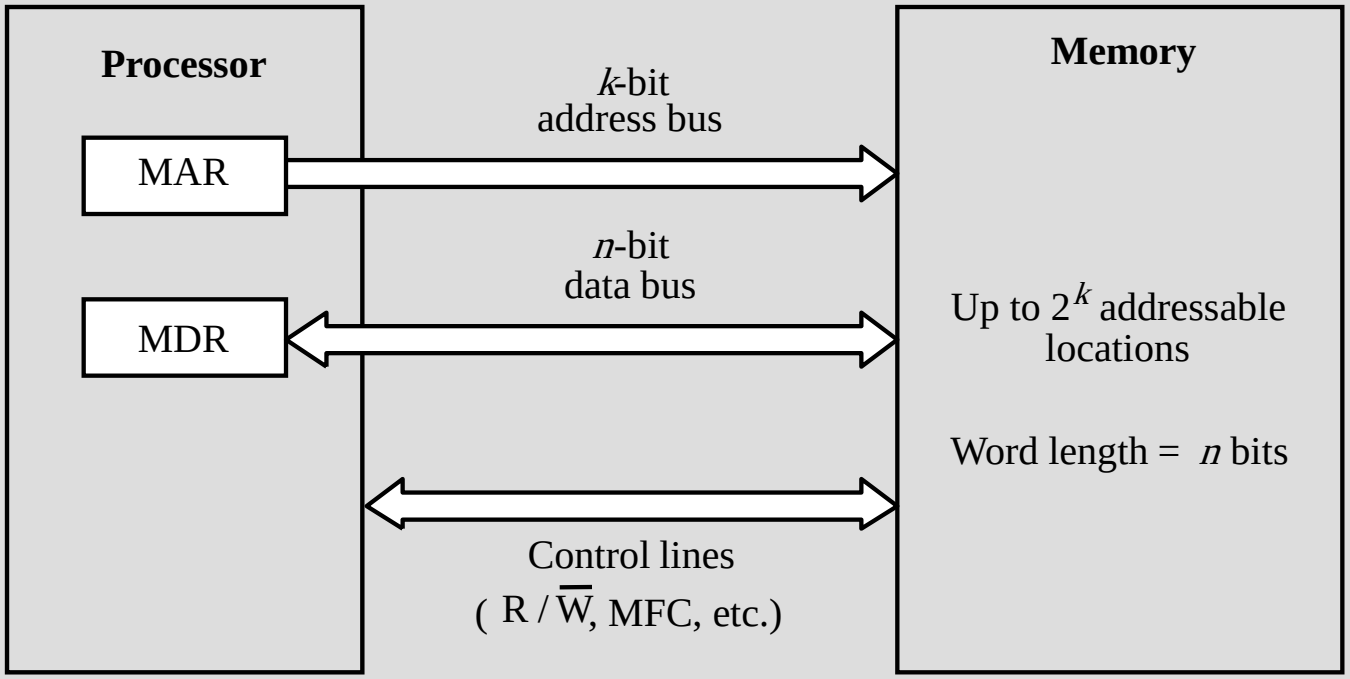
Lecture 6(part 1)
Chapter5

Topics covered:
Memory subsystem

Basic concepts

- Maximum size of the memory depends on the addressing scheme:
  - 16-bit computer generates 16-bit addresses and can address up to 2^{16} memory locations.
  - Number of locations represents the size of the address space of a computer.
- Most modern computers are byte-addressable.
- Memory is designed to store and retrieve data in word-length quantities.
  - Word length of a computer is commonly defined as the number of bits stored and retrieved in one memory operation.

Basic concepts (contd..)



Recall that the data transfers between a processor and memory involves two registers MAR and MDR.
If the address bus is k-bits, then the length of MAR is k bits.
If the word length is n-bits, then the length of MDR is n bits.
Control lines include R/$\overline{W}$ and MFC.
For Read operation R/$\overline{W}$ = 1 and for Write operation R/$\overline{W}$ = 0.

Basic concepts (contd..)

- Measures for the speed of a memory:
  - Elapsed time between the initiation of an operation and the completion of an operation is the memory access time (e.g. the time between the Read & MFC signals).
  - Minimum time between the initiation of two successive memory operations is memory cycle time (e.g. the time between two successive Read operations)
  - Memory Cycle time is slightly longer than memory access time
- In general, the faster a memory system, the costlier it is and the smaller(capacity) it is.
- An important design issue is to provide a computer system with as large and fast a memory as possible, within a given cost target.
- Several techniques to increase the effective size and speed of the memory:
  - Cache memory (to increase the effective speed).
  - Virtual memory (to increase the effective size).

Semiconductor RAM memories

- Random Access Memory (RAM) memory unit is a unit where any location can be accessed in a fixed amount of time, independent of the location's address.
- Internal organization of memory chips:
  - Each memory cell can hold one bit of information.
  - Memory cells are organized in the form of an array.
  - One row is one memory word.
  - All cells of a row are connected to a common line, known as the "word line".
  - Word line is connected to the address decoder.
  - Sense/write circuits are connected to the data input/output lines of the memory chip.

Semiconductor RAM memories (contd..)

- Static RAMs (SRAMs): –volatile memory
  - Consist of circuits that are capable of retaining their state as long as the power is applied.
  - Volatile memories, because their contents are lost when power is interrupted.
  - Access times of static RAMs are in the range of few nanoseconds.
  - Capacity is low (each cell consists of 6 transistors)
  - However, the cost is usually high.
- Dynamic RAMs (DRAMs):- volatile memory
  - Do not retain their state indefinitely.
  - Contents must be periodically refreshed.
  - Contents may be refreshed while accessing them for reading.
  - Access time is longer than SRAM
  - Capacity is higher than SRAM (each cell consists of 1 transistor)
  - Cost is lower than SRAM

Static and Dynamic Memory Cells

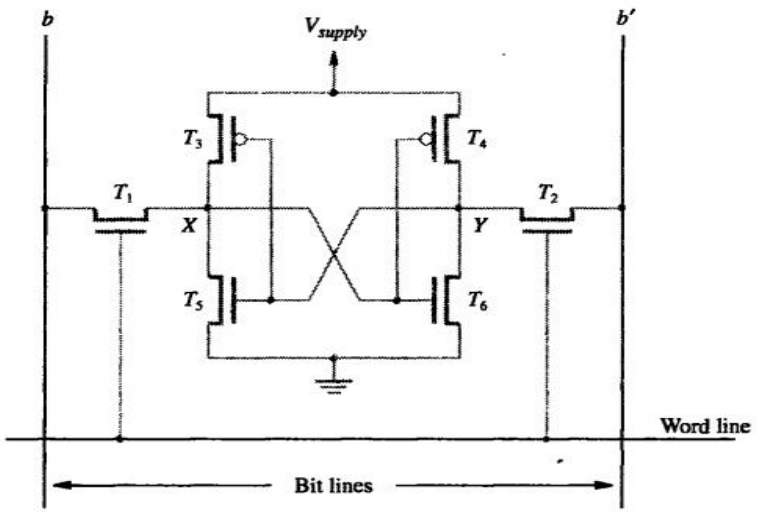


Figure 5.5 An example of a CMOS memory cell.

Static RAM cell

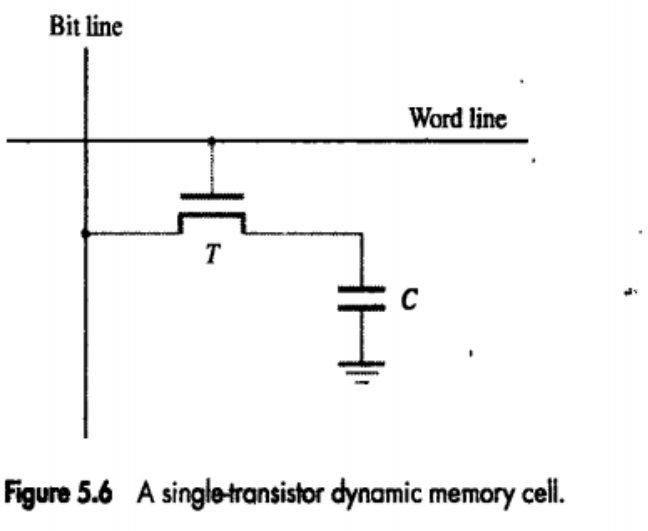
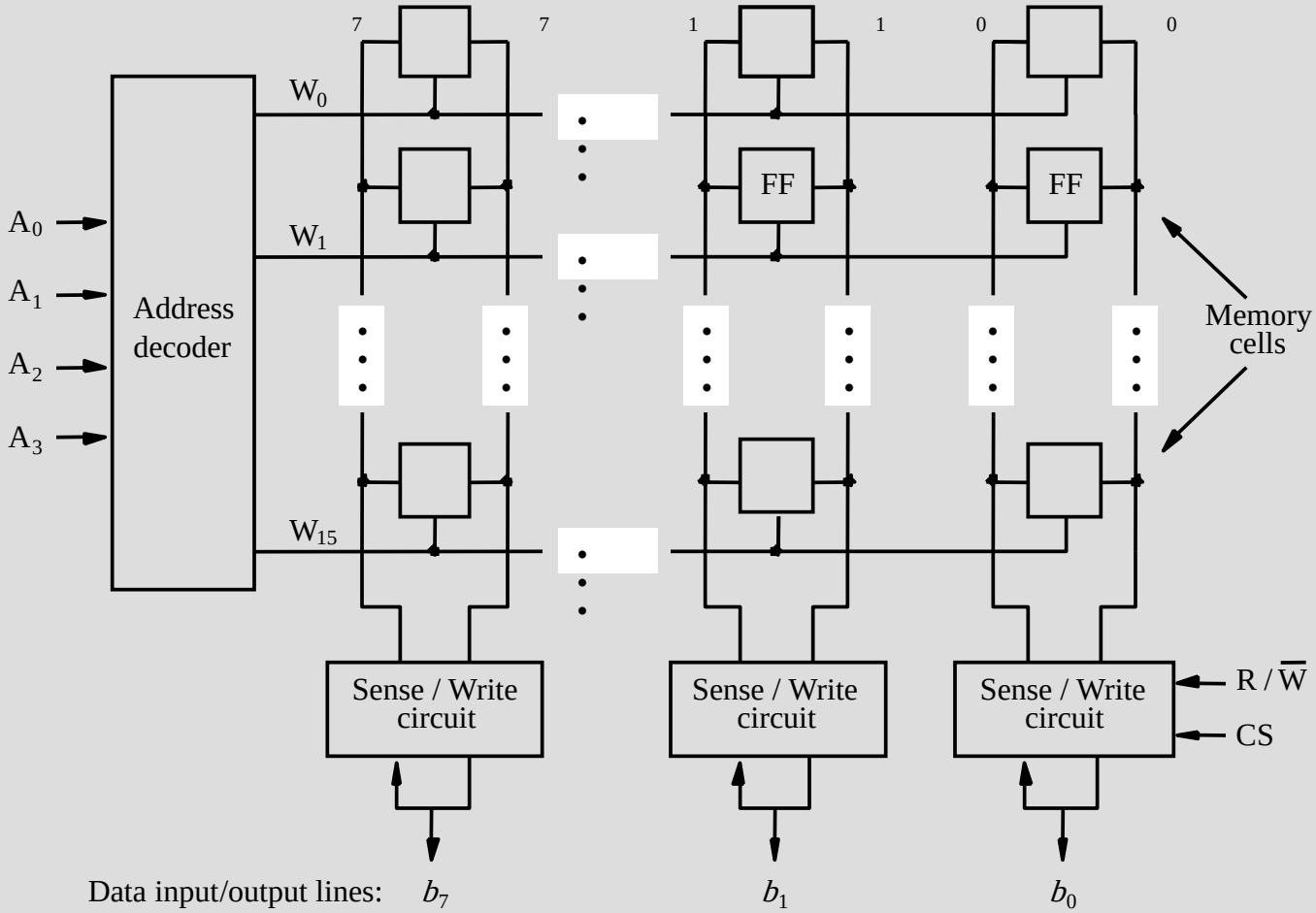


Figure 5.6 A single-transistor dynamic memory cell.

Dynamic RAM cell

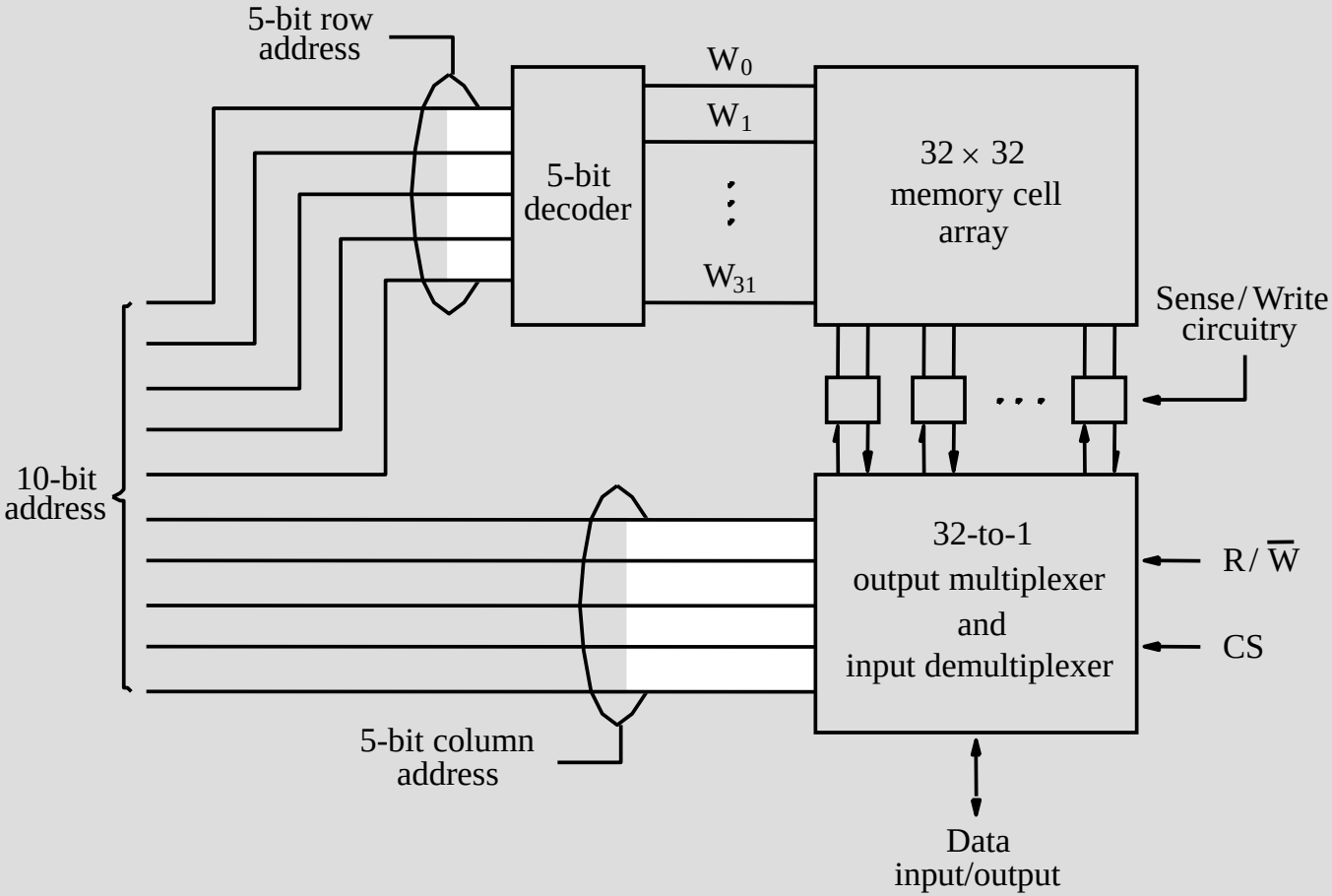
Internal Organization of Memory Chips

Internal organization of memory chips size=128 bits (16x8)



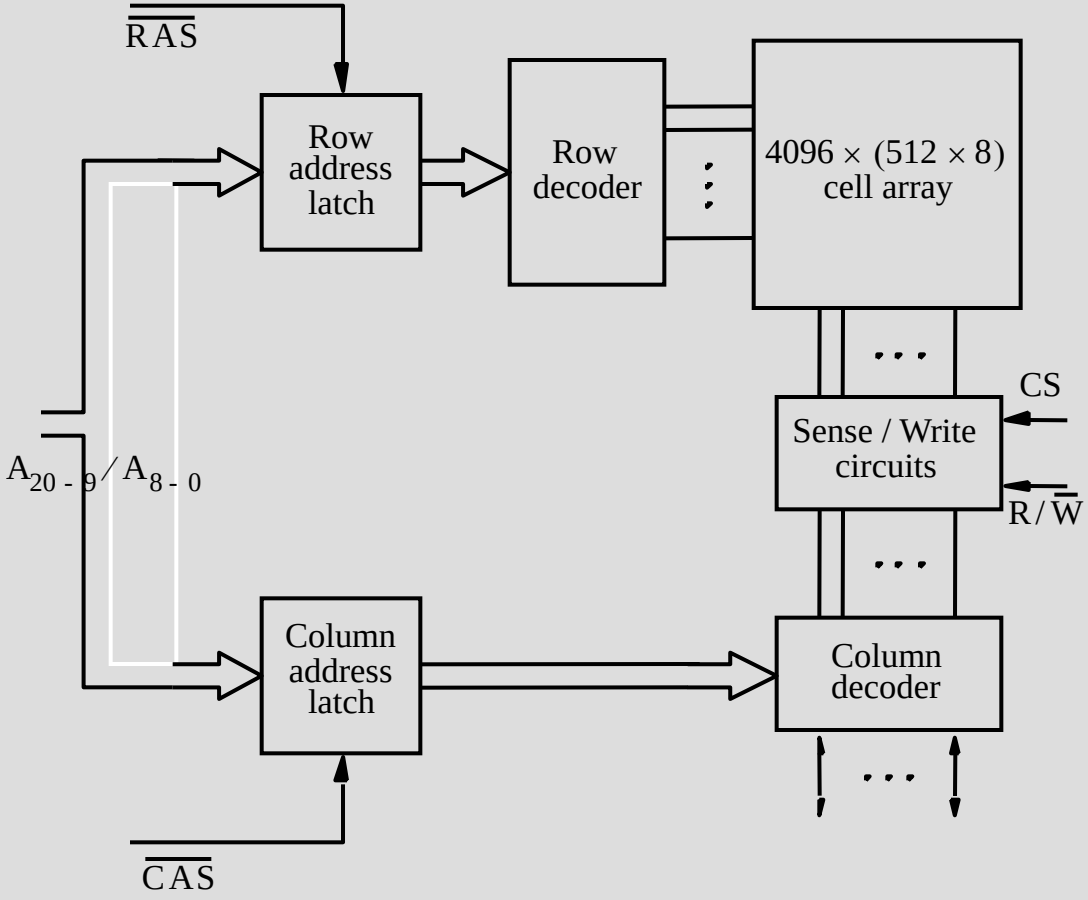
Internal Organization of Memory Chips

Internal organization of memory chips size=1k bits (1024x1)



Asynchronous DRAM

Internal organization of a Dynamic RAM memory chip (16M= 2Mx8)



- *Organized as 4kx4k array. 4096 cells in each row are divided into 512 groups of 8.*
- *Each row can store 512 bytes. 12 bits to select a row, and 9 bits to select a group of 8 bits in a row.*
- *Total of 21 bits. (2 MB)*
- *Reduce the number of bits by multiplexing row and column addresses.*
- *First apply the row address, $\overline{RAS}$ signal latches the row address.*
- *Then apply the column address, $\overline{CAS}$ signal latches the address.*
- *Timing of the memory unit is controlled by a specialized unit which generates $\overline{RAS}$ and $\overline{CAS}$.*
- *This is **asynchronous** DRAM.*

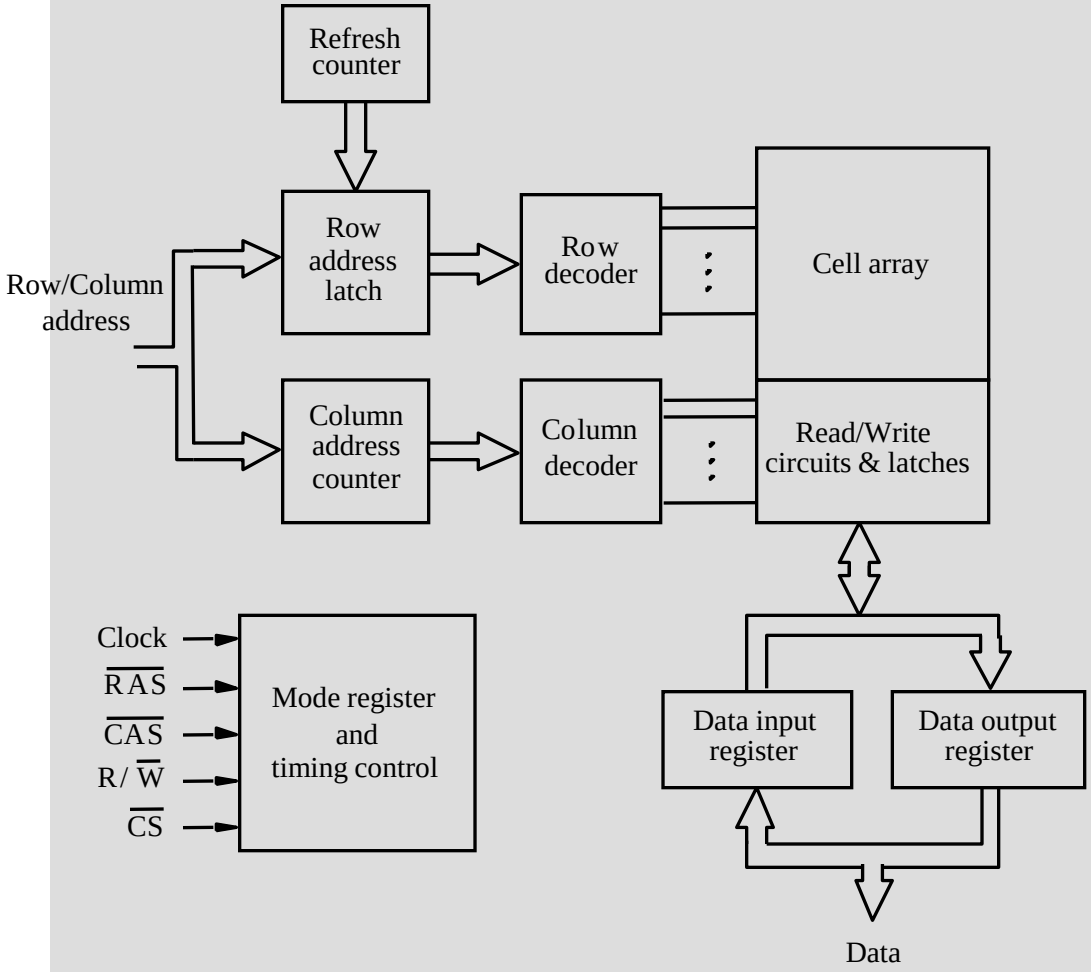
Fast Page Mode

- Recall the operation of the memory:
  - First all the contents of a row are selected based on a row address.
  - Particular byte is selected based on the column address.
- Suppose if we want to access the consecutive bytes in the selected row.
- This can be done without having to reselect the row.
  - Add a latch at the output of the sense circuits in each row.
  - All the latches are loaded when the row is selected.
  - Different column addresses can be applied to select and place different bytes on the data lines.

Fast Page Mode

- Consecutive sequence of column addresses can be applied under the control signal CAS, without reselecting the row.
  - Allows a block of data to be transferred at a much faster rate than random accesses.
  - A small collection/group of bytes is usually referred to as a block.
- This transfer capability is referred to as the fast page mode feature.
- Page : a larger groups of bytes.(Large blocks of data)

Synchronous DRAM(SDRAM)



- *Operation is directly synchronized with processor clock signal.*
- ***Synchronous DRAMs.***
- *The outputs of the sense circuits are connected to a **latch**.*
- *During a **Read** operation, the contents of the cells in a row are loaded onto the latches.*
- *During a **refresh** operation, the contents of the cells are refreshed without changing the contents of the latches.*
- *Data held in the latches correspond to the selected columns are transferred to the output.*
- *For a **burst mode** of operation, successive columns are selected using column address counter and clock. $\overline{CAS}$ signal need not be generated externally.*

Synchronous DRAM(SDRAM)

SDRAMs have several different modes of operation, which can be selected by writing control information into a *mode* register. For example, burst operations of different lengths can be specified. The burst operations use the block transfer capability described above as the fast page mode feature. In SDRAMs, it is not necessary to provide externally generated pulses on the CAS line to select successive columns. The necessary control signals are provided internally using a column counter and the clock signal. New data can be placed on the data lines in each clock cycle. All actions are triggered by the rising edge of the clock.

Burst Read Operation

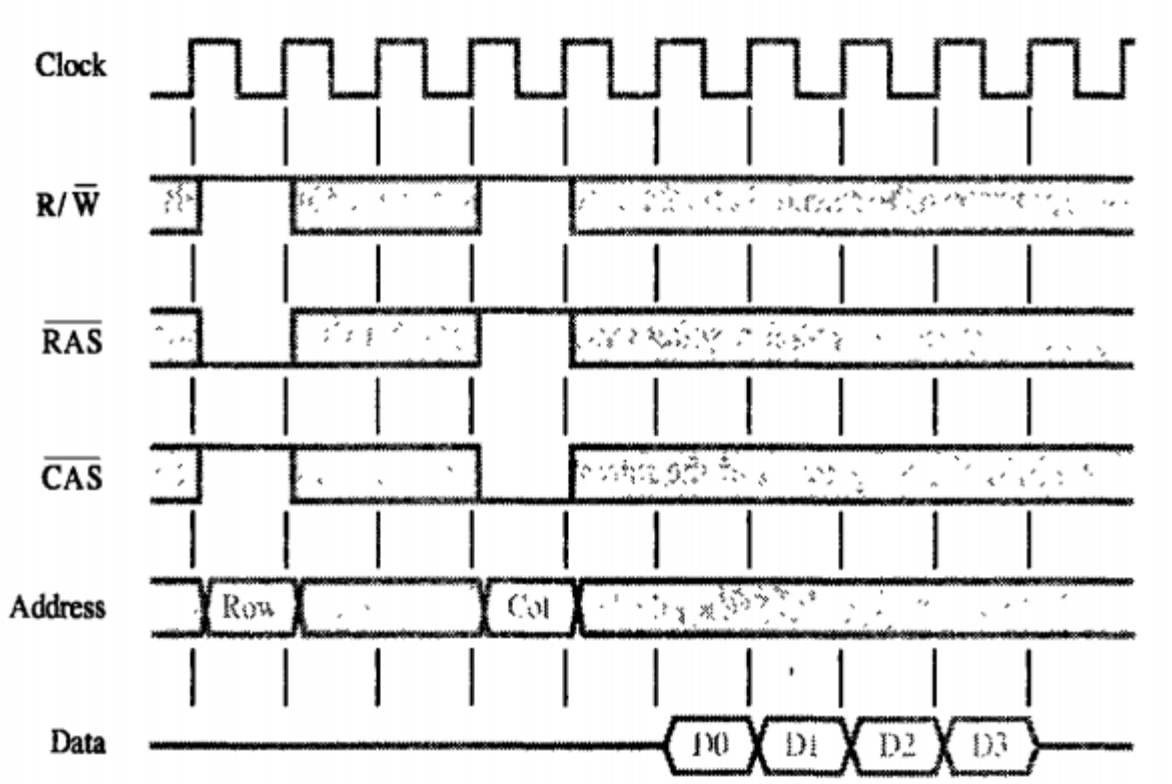
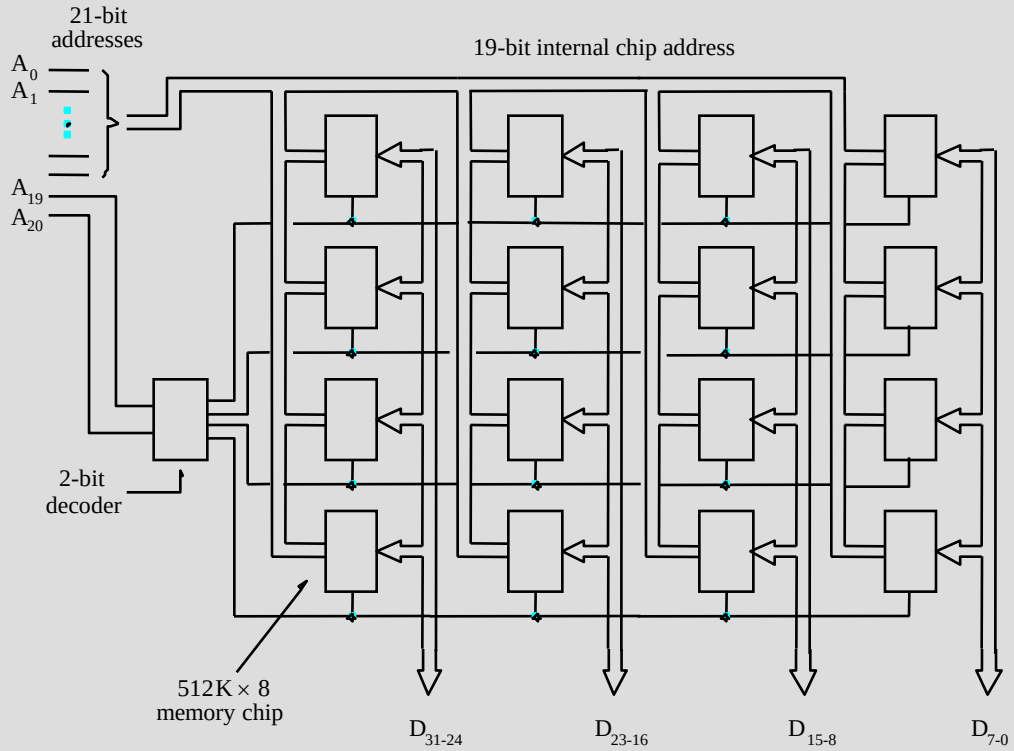
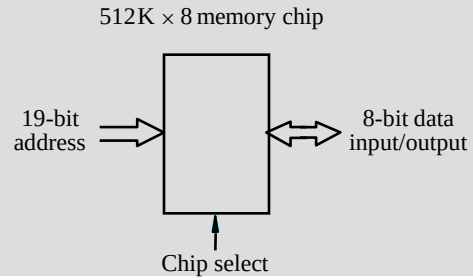


Figure 5.9 Burst read of length 4 in an SDRAM.

Large RAM Design Example





Implement a memory unit of 2M words of 32 bits each.
Use 512Kx8 static memory chips.
Each column consists of 4 chips.
Each chip implements one byte position.
A chip is selected by setting its chip select control line to 1.
Selected chip places its data on the data output line, outputs of other chips are in high impedance state.
***21** bits to address a 32-bit word.*
*High order **2** bits are needed to **select** the **row**, by activating the four Chip Select signals.*
***19** bits are used to access specific **byte** locations inside the **selected chip**.*

Semiconductor RAM memories (contd..)

- Large dynamic memory systems can be implemented using DRAM chips in a similar way to static memory systems.
- Placing large memory systems directly on the motherboard will occupy a large amount of space.
  - Also, this arrangement is inflexible since the memory system cannot be expanded easily.
- Packaging considerations have led to the development of larger memory units known as SIMMs (Single In-line Memory Modules) and DIMMs (Dual In-line Memory Modules).
- Memory modules are an assembly of memory chips on a small board that plugs vertically onto a single socket on the motherboard.
  - Occupy less space on the motherboard.
  - Allows for easy expansion by replacement.

SDRAM Refresh Time

- Recall that the rows of the Dynamic RAM need to be accessed periodically in order to be refreshed.
- Older DRAMs typical refreshing period was 16 ms.
- SDRAMs typical refreshing period is 64 ms.
- Let us consider a SDRAM with 8192 rows:
  - Suppose it takes 4 clock cycles to access each row.
  - Total of 4x8192 = 32767 clock cycles.
  - If the clock rate is 133 MHz, then the total refresh time is 0.000246 seconds. = (32767 * 1/133 M)

Memory Performance(Latency and Bandwidth)

- Data is transferred between the processor and the memory in units of single word or a block.
- Speed and efficiency of data transfer has a large impact on the performance of a computer.
- Two metrics of performance: Latency and Bandwidth.
- **Latency**:
  - **Time taken to transfer a single word of data to or from memory**.
  - Definition is clear if the memory operation involves transfer of a single word of data.
  - In case of a block transfer, latency is the time it takes to transfer first word of data.
  - Time required to transfer first word in a block is substantially larger than the time required to transfer consecutive words in a block.

Latency and Bandwidth

- How much time is needed to transfer a single block of data.
- Blocks can be variable in size, it is useful to define a performance measure in terms of **the number of bits or bytes transferred in one second.**
- This performance measure is referred to as memory **bandwidth**.
- Bandwidth of a memory unit depends on:
  - **Speed** of access to the chip.
  - How many **bits** can be accessed in parallel.
- Bandwidth of data transfer between processor and memory unit also depends on:
  - Transfer capability of the links that connect the processor and memory, or the speed of the bus.

Double Data Rate SDRAM

- Standard SDRAM perform its operations on the rising edge of the clock signal.
- DDR SDRAMs can transfer data on both edges of the clock so its bandwidth is doubled while its latency is the same as the standard SDRAM

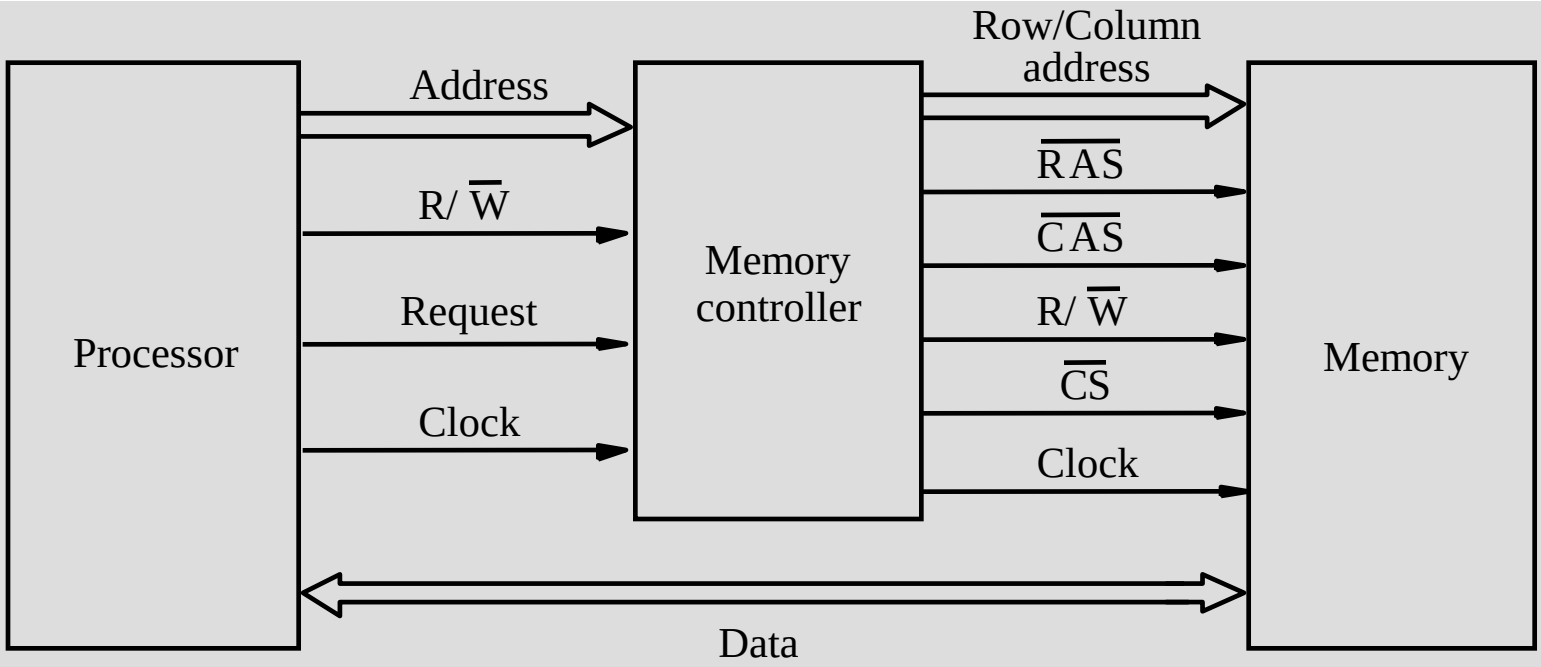
Memory systems

- Various **factors** such as cost, speed, power consumption and size of the chip determine how a **RAM** is chosen for a given application.
- **Static RAMs**:
  - Chosen when **speed** is the primary concern.
  - **Circuit** implementing the basic cell is highly **complex**, so **cost** and **size** are affected.(use 6 transistors for each cell)
  - Used mostly in **cache** memories.
- **Dynamic RAMs**:
  - Predominantly used for implementing computer **main** memories.
  - **High densities** available in these chips(use one transistor for each cell).
  - **Economically** viable for implementing large memories.

Memory controller

- Recall that in a dynamic memory chip, to reduce the number of pins, multiplexed addresses are used.
- Address is divided into two parts:
  - High-order address bits select a row in the array.
  - They are provided first, and latched using RAS signal.
  - Low-order address bits select a column in the row.
  - They are provided later, and latched using CAS signal.
- However, a processor issues all address bits at the same time.
- In order to achieve the multiplexing, memory controller circuit is inserted between the processor and memory.

Memory controller (contd..)



Memory controller accepts the complete address, $R/\overline{W}$ signal from the processor, under the request of a control signal.

Controller forwards row and column address portions to the memory, and issues $\overline{RAS}$ and $\overline{CAS}$ signals.

It also sends $R/\overline{W}$ and $\overline{CS}$ signals to the memory.

Data lines are connected directly between the processor and memory.

Read-Only Memories (ROMs)

- SRAM and SDRAM chips are volatile:
  - Lose the contents when the power is turned off.
- Many applications need memory devices to retain contents after the power is turned off.
  - For example, computer is turned on, the operating system must be loaded from the disk into the memory.
  - Store instructions which would load the OS from the disk.
  - Need to store these instructions so that they will not be lost after the power is turned off.
  - We need to store the instructions into a non-volatile memory.
- Non-volatile memory is read in the same manner as volatile memory.
  - Separate writing process is needed to place information in this memory.
  - Normal operation involves only reading of data, this type of memory is called Read-Only memory (**ROM**).

ROM Cell

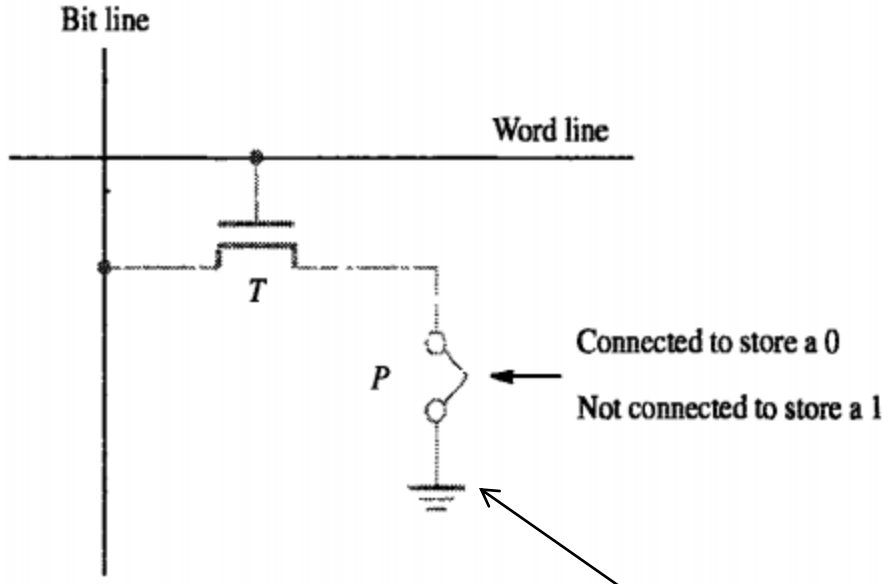


Figure 5.12 A ROM cell.

Read-Only Memories (contd..)

- Read-Only Memory:
  - Data are written into a ROM when it is manufactured.
- Programmable Read-Only Memory (PROM):
  - Allow the data to be loaded by a user.
  - Process of inserting the data is irreversible.
  - Storing information specific to a user in a ROM is expensive.
  - Providing programming capability to a user may be better.
- Erasable Programmable Read-Only Memory (EPROM):
  - Stored data to be erased and new data to be loaded.
  - Flexibility, useful during the development phase of digital systems.
  - Erasable, reprogrammable ROM.
  - Erasure requires exposing the ROM to UV light.

Read-Only Memories (contd..)

- Electrically Erasable Programmable Read-Only Memory (EEPROM):
  - To erase the contents of EPROMs, they have to be exposed to ultraviolet light and Physically removed from the circuit.
  - Do not have to be removed for erasure.
  - In EEPROMs the contents can be stored and erased **electrically**.
  - It is possible to erase the contents of the cell selectively.
  - Disadvantage: different voltages are needed for erasing, writing and reading the stored data.

Flash memory

- Flash memory has similar approach to EEPROM.
- Read the contents of a single cell, but write the contents of an entire block of cells.
- Flash devices have greater density.
  - Higher capacity and low storage cost per bit.
- Power consumption of flash memory is very low, making it attractive for use in equipment that is battery-driven.
- Single flash chips are not sufficiently large, so larger memory modules are implemented using flash cards and flash drives.
- Disadvantages: in respect to hard disks

- smaller capacity and higher cost per bit.
- will deteriorate after it has been written a number of times(short life time).

Acronyms

RAM --Random Access Memory time taken to access any arbitrary location in memory is constant (c.f., disks)

ROM --Read Only Memory ROMs are RAMs which can only be written to once; thereafter they can only be read
Older uses included storage of bootstrap info

PROM --Programmable ROM A ROM which can be bench programmed

EPROM --Erasable PROM A PROM which can be erased for rewriting

EEPROM --Electrically EPROM A PROM which can be erased electrically.

SRAM --Static RAM RAM chip which loses contents upon power off

DRAM --Dynamic RAM RAM chip whose contents need to be refreshed.

SDRAM --Synchronous DRAM RAM whose memory operations are synchronized with a clock signal.